**CSCI446/946: Big Data Analytics**

**Assignment One (10%)**

Student Name: Group – 17 led by Karan Goel

Student ID:

Mark out of 10: 2.5 + 1.75 + 1.50 + 1.50 = 7.25

**Task 1: Problem Analysis and Data Preprocess (4 marks)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Criterion | Marks | Comment |
| 1 | Task 1 is answered in two sections in the report as “Problem Analysis” and “Data Preprocess” **(0.5 mark)** | 0.5 | Well answered. |
| 2 | “*Why would you choose all or part of data from the NewChic dataset*” is answered with explanation. **(1 mark)** | 0.25 | Should explain more the reason of using NewChic dataset. |
| 3 | “*How would you define top 10 and the best*” is answered with explanation. **(1 mark)** | 0.25 | Should provide how would you define top 10 and the best and explain the reason. |
| 4 | “*Combining CSVs in one sheet and report: matched, removed columns*” is answered with explanation. **(1 mark)** | 1 | Well answered. |
| 5 | The code of Task 1 is submitted and of good quality. **(0.5 mark)** | 0.5 | Overall good |
|  | Total marks **(4 marks)** | 2.5 |  |

**Task 2: Clustering (2 marks)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Criterion | Marks | Comment |
| 1 | At least two clustering algorithms are applied, with justification and explanation. **(0.5 mark)** | 0.5 | Hierarchical and K-means are used |
| 2 | Detailed explanations of the steps of each algorithm **(0.5 mark)** | 0.25 | Steps clustering methods are not clearly described and explained, just provided pseudo-code. |
| 3 | Result of each algorithm in a table **(0.5 mark)** | 0.5 | Results of each algorithm shown in a table. Illustrations need to be improved. |
| 4 | The code of Task 2 is submitted and of good quality. **(0.5 mark)** | 0.5 | Overall good |
|  | Total marks **(2 marks)** | 1.75 |  |

**Task 3: Classification (2 marks)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Criterion | Marks | Comment |
| 1 | At least two classification algorithms are applied, with justification and explanation. **(0.5 mark)** | 0.5 | Decision Tree and KNN Classifier are used |
| 2 | Detailed explanations of the steps of each algorithm **(0.5 mark)** | 0.25 | Advantage and disadvantages are discussed, but steps are not clearly described and explained, just provided code. |
| 3 | Result of each algorithm in a table **(0.5 mark)** | 0.25 | Should clearly explain the results in the table. |
| 4 | The code of Task 3 is submitted and of good quality. **(0.5 mark)** | 0.5 | Overall good |
|  | Total marks **(2 marks)** | 1.50 |  |

**Task 4: Result Discussion (2 marks)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Criterion | Marks | Comment |
| 1 | Report the 10 best products **(0.5 mark)** | 0.5 | Reported |
| 2 | Report the best category **(0.5 mark)** | 0.5 | Reported |
| 3 | Provide suggestions to NewChic **(0.5 mark)** | 0 | Should provide suggestions to NewChic |
| 4 | Detailed discussion is provided **(0.5 mark)** | 0.5 | Good |
|  | Total marks **(2 marks)** | 1.50 |  |